Challenges in Systematic Reviews of Educational Intervention Studies

Darcy Reed, MD; Eboni G. Price, MD, MPH; Donna M. Windish, MD, MPH; Scott M. Wright, MD; Aysegul Gozu, MD; Edbert B. Hsu, MD, MPH; Mary Catherine Beach, MD, MPH; David Kern, MD, MPH; and Eric B. Bass, MD, MPH

Educators have recognized the need to apply evidence-based approaches to medical training. To do so, medical educators must have access to reliable evidence on the impact of educational interventions. This paper describes 5 methodologic challenges to performing systematic reviews of educational interventions for health care professionals: finding reports of medical education interventions, assessing quality of study designs, assessing the scope of interventions, assessing the evaluation of interventions, and synthesizing the results of educational interventions. We offer suggestions for addressing these challenges and make recommendations for reporting, reviewing, and appraising interventions in medical education.


For author affiliations, see end of text.

In recent years, educators have increasingly emphasized the need to apply evidence-based approaches to medical training (1, 2). To do so, medical educators must have access to the best available evidence on the impact of educational interventions. Although the evidence base for medical education interventions is more limited than the evidence base for clinical interventions, it is growing as demand for evidence and outcomes research in medical education increases (3–5).

Educators increasingly recognize the important contribution of systematic reviews to medical education (6). Several systematic reviews of interventions in medical education have been published recently (7–10), and many systematic reviews have been published in nonmedical education (11–14) and patient education (15–20) literature. Systematic reviews will have an important role in synthesizing the growing body of evidence in medical education, but unique methodologic challenges need to be addressed.

In this paper, we describe methodologic challenges likely to be encountered when conducting a systematic review of interventions in medical education, identify limitations in the methods that have been used to assess medical education interventions, and provide recommendations for reporting and reviewing studies of interventions in medical education.

Review of Published Guides for Reports of Interventions in Medical Education

Since systematic reviews depend on the quality of original reports about specific educational interventions, we first identified published guides for conducting, evaluating, and reporting educational interventions in medicine (1, 21–24). Each guide contains a structured approach to the appraisal of educational interventions. Three of the reports were developed by consensus opinion of international experts in medical education (1, 21, 22). The remaining guides represent opinions of medical educators with expertise in development, evaluation, and appraisal of curricula (23, 24). We extracted recommendations from these guides and prepared Table 1 to show the similarities among them. We then formulated specific questions to consider when appraising reports of educational interventions (last column of Table 1).

Challenge: Identifying Reports on the Effectiveness of Interventions in Medical Education

Finding reports for systematic review requires identifying all relevant sources of studies and executing a comprehensive search strategy. Both tasks are uniquely challenging to reviewers of educational interventions because no single database is devoted to medical education. Reviewers should consider using multiple databases that include reports from the broad field of education (25). In addition to MEDLINE or PubMed, databases containing educational interventions include the Educational Resource Information Center (ERIC), British Education Index (BEI), PsycINFO, and the Cumulative Index to Nursing and Allied Health Literature (CINAHL) (1). The Campbell Collaboration (www.campbellcollaboration.org) maintains a database of trials in the Social, Psychological, Education and Criminological Trials Registry (C2-SPECTR) and prepares systematic reviews of educational interventions that are reported in the Register of Interventions and Policy Evaluation (C2-RIPE) database.

Reviewers may find additional reports on educational interventions and their effectiveness in the databases mentioned above. The importance of these resources cannot be overstated, as they provide a comprehensive and systematic approach to identifying and evaluating educational interventions in medical education.
Table 1. Published Guides and Recommended Questions for Appraising Reports of Medical Education Interventions

<table>
<thead>
<tr>
<th>Variable</th>
<th>Published Guides, Year (Reference)</th>
<th>Recommended Questions for Appraising Reports of Medical Education Interventions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question</td>
<td>Clearly stated</td>
<td>–</td>
</tr>
<tr>
<td>Rationale</td>
<td>Explicit Identified learning need</td>
<td>Explicit Founded on theory Adequate literature review</td>
</tr>
<tr>
<td>Objectives</td>
<td>Specific, observable, and achievable</td>
<td>Specific, clearly stated</td>
</tr>
<tr>
<td>Study design</td>
<td>Design appropriate for question Methods appropriate to measure outcomes</td>
<td>Appropriate design for question Detailed description of methods to allow reproducibility Description of recruitment methods</td>
</tr>
<tr>
<td>Intervention</td>
<td>Description of structure, process, and content Description of educational context and learners</td>
<td>Description of educational context, stage of learners, and details of program</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Appropriate outcomes selected Reliability and validity of outcomes considered</td>
<td>Planned in advance Detailed description Linked to research question and objectives Appropriate statistical analysis</td>
</tr>
<tr>
<td>Results</td>
<td>Alternate explanation of results Explanation of unanticipated outcomes</td>
<td>Significance of results Conclusions and recommendations linked to and justified by results</td>
</tr>
</tbody>
</table>

References: Morrison et al., 1999 (21) Harden, 1999 (1) Green, 2001 (23) Kern et al., 1998 (24)
interventions by searching the Internet. An efficient approach is to perform a targeted search of Web sites of medical education organizations and professional societies such as the Association of American Medical Colleges (www.aamc.org), the Accreditation Council for Graduate Medical Education (www.acgme.org), and the American Medical Association (www.ama-assn.org) (26).

Additional reports can often be found by reviewing manuscript reference lists of pertinent articles, searching citation indices such as the International Scientific Index Web of Knowledge, hand searching key educational journals, and consulting experts in medical education.

In developing a search strategy, selection of search terms requires careful thought because concepts, subject headings, and keywords vary among databases. An iterative approach using several related terms is often required when an unfamiliar database is being used or when a new topic is being researched. Furthermore, some topics in medical education do not map well to medical subject headings and keywords. For example, when searching various databases for articles about “feedback,” the Best Evidence Medical Education Group FEENASS (feedback in assessment) found a sensitivity of 6.5% to 19.6% and a specificity of 17.5% compared with hand searching (27). Reviewers can increase the sensitivity and specificity of their searches by becoming familiar with subject headings used in individual databases.

The Medical Subject Heading (MeSH) browser at the National Library of Medicine’s Web site (www.nlm.nih.gov/mesh/MBrowser.html) is a useful tool with which to identify MeSH terms relevant to a given topic in MEDLINE. For example, entering “medical education” into the MeSH browser yields the main MeSH headings of education, medical, continuing; education, medical, graduate; education medical, undergraduate; and internship and residency.

CHALLENGE: ASSESSING THE QUALITY OF STUDY DESIGNS
Lack of Randomized, Controlled Trials

In the hierarchy of study designs, the randomized, controlled trial remains the gold standard. The process of randomization minimizes bias and increases the likelihood that groups will be similar at baseline. With equivalent groups, a controlled intervention, and effective assessment tools, differences in outcome can more readily be attributed to the intervention and not to confounding factors. When a systematic review is performed only on studies that have used the strongest study design, evidence can be combined by using established methods for synthesizing results of randomized, controlled trials of clinical interventions.

Despite the methodologic strengths of randomized, controlled trials, few educational interventions use this design. However, there are notable exceptions (28–31). Frequently, educational researchers rely on observational studies and quasi-experimental designs. Many issues may be contributing to the paucity of randomized trials in medical education. The first involves resources. To carry out a randomized trial, researchers must have support from an institution for both faculty time and structured evaluative strategies (32). The dearth of funding for medical education research also limits educators’ ability to conduct more rigorous studies (33). Second, educators continue to debate the ethics of randomly assigning learners to receive or not to receive an intervention (2, 34).

Heterogeneity in Study Designs

Reviewers ideally should focus on studies that use similar designs with similar interventions and control groups. In addition to randomized, controlled trials, reviewers should consider including studies that match for baseline characteristics between groups. This method can provide some equivalency between groups by controlling for selected confounding factors. Without randomization or matched comparison groups, bias and confounding are more likely to influence the observed results. Selection bias may contribute to differences in outcomes among groups. In studies using historical controls, differences in data collection may make it difficult to determine whether groups are equivalent at baseline.

Studies using less methodologic rigor present a challenge to reviewers in their synthesis of evidence. For example, in a systematic review of the effectiveness of hospital disaster drills, reviewers found studies that ranged from simple descriptions of institutional exercises without evaluation data to large regional exercises with extensive evaluation results (35). Most of the evaluations lacked an appropriate comparison group. The broad array of study designs, infrequent use of comparison groups, and variable rigor left the reviewers with insufficient evidence to form firm conclusions about the effectiveness of methods for training hospital staff in disaster preparedness.

The Research Triangle Institute—University of North Carolina Evidence-based Practice Center also reported great heterogeneity among studies of community-based participatory research, many of which involved educational interventions (36). Of 60 studies addressing implementation and outcomes of community-based participatory research, only 12 exhaustively evaluated the intervention. Among these, 4 were randomized trials, 5 were quasi-experimental, and 3 used nonexperimental designs. Heterogeneity of study designs prohibited quantitative synthesis of the impact of community-based participatory research on health outcomes.

Other Study Design Limitations

Many educational studies are designed to involve a single institution. This may limit their ability to achieve statistical power and generalizability. Unfortunately, education researchers often lack the resources to conduct multi-institutional studies. Furthermore, few studies are designed to demonstrate a link between the educational intervention and a clinical outcome, despite an identified
need for clinical outcomes research in medical education (3, 4, 37).

**Measuring Study Quality**

Evaluating the quality of studies about educational interventions is a complex task. Reviewers may attempt to assess the quality of each study included in a review and therein assess the strength of the published evidence.

To address the challenge of assessing study quality in systematic reviews, the Research Triangle Institute—University of North Carolina Evidence-based Practice Center prepared a report on systems for rating the strength of scientific evidence (38). The report details 3 domains to consider in grading evidence in the literature: quality, quantity, and consistency. For quality, an aggregate rating may be determined for each of the individual studies based on multiple domains, including the study question, interventions, outcomes, validity, and data analysis. The quantity measure is used to assess the magnitude of the effect observed based on the number of studies in the review and their power or sample size. Consistency refers to the extent to which similar findings are reported independently of the type of study. From these 3 categories, an overall assessment of the evidence can be determined.

Although the recommendations for assessing study quality generally have been used in systematic reviews of clinical interventions, the recommendations also can be applied to reviews of educational interventions. For example, the Johns Hopkins Evidence-based Practice Center used the recommendations to assess the strength of evidence on the effectiveness of cultural competence interventions for improving the quality of health care for racial and ethnic minorities (39). The reviewers developed a quality grading scale based on the number of relevant controlled trials and the percentage of studies that used objective methods to assess outcomes. Because a limited number of studies used clearly defined comparison groups, the Evidence-based Practice Center team assigned low grades to the body of evidence on the key questions of the review.

Assessing quality domains using a numerical scale is another potential approach to evaluating study quality. Beach and colleagues assessed study quality of cultural competence curricula in 5 domains: representativeness of targeted health care providers and/or patients, potential for bias and confounding, description of interventions, outcomes assessment, and analytic approach (39). The instrument was adapted from instruments used to assess the quality of studies of clinical interventions, and it incorporated fundamental principles of curriculum development and evaluation (24). Two reviewers independently applied the instrument to each study. Interobserver reliability of the instrument was excellent (40).

Reviewers should note that using study quality assessments in reviews of educational interventions is likely to have the same limitations as in reviews of clinical interventions. For example, differences in quality of reporting may obscure important differences in methodologic quality. This poses a problem for reviews of educational interventions because of the tremendous heterogeneity in how results are reported. Juni and colleagues found that the use of different quality scoring scales led to different assessments of the quality of studies and sometimes changed the estimated effect of a clinical intervention (41). Reviewers should be aware that they can introduce bias into the quality assessment when attempting to weigh the evidence. Extensive rater training can help improve inter-rater reliability. Although educators may differ about how to assess the quality of educational studies, the similarities in the guides summarized in Table 1 suggest that a consensus may be emerging.

**CHALLENGE: ASSESSING THE SCOPE OF INTERVENTIONS IN MEDICAL EDUCATION**

**Incomplete Descriptions of Educational Interventions**

Guides for appraising reports of educational interventions recommend assessing whether authors completely describe their intervention (Table 1). However, many reviews of educational interventions have found limited descriptions of interventions. For example, in a systematic review of resident research curricula, Hebert and colleagues found that most studies did not adequately report on needs assessment, curriculum development, learning objectives, instructional strategies, evaluation methods, or challenges encountered during implementation of the interventions (8). Berkman and colleagues noted similar limitations in their review of literacy and health outcomes (42). Berkman and colleagues’ synthesis of the literature was limited by poor descriptions of interventions and lack of reporting of methods for outcome assessments (42). Cauffman and colleagues also identified incomplete reporting among randomized, controlled trials of continuing medical education interventions (43). To obtain the missing information, authors of each report were interviewed. The study by Cauffman and colleagues illustrates that contacting study authors is 1 potential solution to gathering the evidence when intervention descriptions are not adequately reported; however, this strategy is labor intensive and generally not feasible.

Reports of educational interventions frequently fail to describe specific learning objectives and do not provide examples of educational content. In a methodologic review of cultural competence curricula, Price and colleagues report that learning objectives, curricular content, and teaching methods are frequently missing from educational reports (44).

One potential reason for incomplete reporting of educational interventions is the word count limitation imposed by journals. Authors publishing curricular papers may deal with this limitation by providing an overview of the curriculum and describing 1 unit or part of the curriculum that exemplifies the content. Alternatively, authors...
may place details about the curriculum into appendices. Some journals may choose to publish these appendices in either the print or electronic versions of the journal. Authors may also refer readers to other sources for details about a curriculum, such as a previous publication or an existing Web site.

**Characterizing Components of Curricular Content and Teaching Strategies**

To achieve educational goals and objectives, a variety of teaching strategies may be used. While this may enhance curricula, it complicates synthesis of the evidence. Beach and colleagues noted problems with synthesizing the evidence for cultural competence training because of the heterogeneity of curricular content and teaching strategies (39). Most interventions addressed multiple content areas and used more than 1 teaching method. No 2 studies used the same teaching strategies. In a systematic review of postgraduate teaching in evidence-based medicine and critical appraisal, Coomarasamy and colleagues reported that the 17 studies eligible for review reported diverse interventions that took various formats (for example, workshops, seminars, and journal clubs) (9). In their review of literacy and numeracy, Berkman and colleagues found that the “use of multimodal components inhibits understanding of which portions produced positive effects” (42). They advocated study designs and analyses that isolate the effects of intervention components to determine “how much intervention is enough” to obtain the desired outcome (42). It is important that researchers thoroughly document details of the educational intervention and report on whether they attempted to standardize the intervention to facilitate reproducibility.

**Intensity of Interventions**

Intervention intensity refers to the frequency, duration, and concentration of the intervention. The intensity of educational interventions varies widely from brief teaching encounters to courses spanning years (39, 42). In a review of the effectiveness of critical appraisal skills training, Taylor and colleagues reported detailed information on intervention lengths (for example, 50 minutes per session for 5 sessions over a 2-month period) and assessed various intervention “dosages” for critical appraisal skills training (10). In performing a systematic review, it is important to assess the intensity of each intervention so that readers can determine the optimal intensity that is both effective and feasible.

**Table 2. Strengths and Limitations of Commonly Used Evaluation Methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Strengths</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rating forms</td>
<td>Economical</td>
<td>Subjective</td>
</tr>
<tr>
<td></td>
<td>Can evaluate almost anything</td>
<td>Rater biases</td>
</tr>
<tr>
<td></td>
<td>Useful for formative evaluation</td>
<td>Inter- and intra-rater reliability</td>
</tr>
<tr>
<td>Self-assessment forms</td>
<td>Economical</td>
<td>Subjective</td>
</tr>
<tr>
<td></td>
<td>Promotes self-assessment</td>
<td>Rater biases</td>
</tr>
<tr>
<td></td>
<td>Useful for formative evaluation</td>
<td>Agreement with objective measurements often low</td>
</tr>
<tr>
<td>Written tests</td>
<td>Objective</td>
<td>Reliability, validity vary</td>
</tr>
<tr>
<td></td>
<td>Widely accepted</td>
<td>Constructing tests resource intensive</td>
</tr>
<tr>
<td>Direct observation</td>
<td>Firsthand data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Immediate feedback to learner</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Observation checklists and training observers increase reliability and validity</td>
<td></td>
</tr>
<tr>
<td>Performance audits</td>
<td>Objective</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unobtrusive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reliability and accuracy enhanced by standardization and training observers</td>
<td></td>
</tr>
</tbody>
</table>

instrument affects the reliability, validity, and feasibility of the evaluation. Using a previously validated instrument is ideal because it will allow comparison of results to previous studies using the same instrument. However, adapting and revalidating an instrument is preferable to using an existing instrument that is not well suited to the study at hand. When an intervention-specific instrument is needed, researchers should seek expert opinion in designing such instruments (24).

The validity and reliability of evaluation instruments used in individual studies should be assessed by researchers performing systematic reviews. However, this can be challenging since many studies in the educational literature fail to describe any validity or reliability testing of the evaluation instruments used. Table 3 provides some common definitions of the types of study reliability and validity to consider.

Gozu and colleagues assessed the reporting of reliability and validity of evaluation instruments in reports of cultural competence training for health professionals (45). Among 34 studies reviewed, 70 unique evaluation instruments were used. Only 17 of the 70 instruments (24%) were validated. Thirteen studies used at least 1 validated instrument. Most studies used evaluation methods without prior validity or reliability testing.

In a review of literacy and health outcomes, Berkman and colleagues assessed the validity and reliability of literacy measurements in the quality assessment of individual articles. They found considerable consistency among evaluation instruments used to assess literacy (42). Of 73 studies, 44 used 1 of 3 instruments to measure literacy. All 3 of these instruments had been validated and were highly correlated with one another (42).

### Challenges in Systematic Reviews of Educational Intervention Studies

#### Synthesizing Results of Interventions in Medical Education

Synthesizing Results of Heterogeneous Studies

Educational interventions occur within many different study settings and involve a wide range of learners. Moreover, studies often report different learning objectives, curricular content, teaching strategies, intervention intensities, study designs, evaluation methods, and measured outcomes. Such heterogeneity complicates synthesis of the evidence and often requires reviewers to make subjective decisions about which aspects of the interventions are most important (46). Nonetheless, heterogeneity may also offer advantages. It allows the reviewer to 1) examine the consistency of findings across studies, settings, and populations as a means of assessing the generalizability of the interventions and 2) assess the relative feasibility and effectiveness of different educational approaches. While no standard guideline exists for how reviewers should integrate heterogeneous evidence, reviewers commonly adopt a qualitative, quantitative, or mixed synthetic approach (46). Thomas and colleagues provide a framework for integrating qualitative and quantitative evidence in systematic reviews using a mixed approach (47).

Qualitative synthesis is a narrative approach in which reviewers organize studies into groups and display results in a manner that helps readers see similarities and differences among studies. Reviewers may decide to construct evidence tables that display detailed information for each study, as was done in the evidence report on cultural competence training of health care professionals (39). An alternative approach is to construct evidence tables that present answers to the reviewer’s key study questions by grouping studies into categories and organizing the content so that readers can look for patterns across groups of studies. In their systematic review of the effectiveness of critical appraisal skills training for clinicians, Taylor and colleagues classified outcomes into broad categories including knowledge, attitudes, skills, and behavior (10). Within each of these categories, they grouped outcomes by positive, negative, or inconclusive results. This allowed the authors to determine which types of outcomes were positively affected by the interventions.

A useful conceptual framework for organizing tables is to show “what works for whom under which circumstances and to what end.” The “what works” may refer to educational theories, learning objectives, interventions, or teaching methods. “Whom” refers to the group of learners targeted by the intervention. The “circumstance” may refer to the intervention setting, duration, and frequency. The

### Table 3. Definitions of Reliability and Validity*

| Reliability: Consistency or reproducibility of measurements | Validity: Degree to which a measurement instrument truly measures what it is intended to measure |
| Intra-rater reliability: Degree to which measurements are the same when repeated by the same person | Face or content validity: Degree to which an instrument accurately represents the skill or characteristic it is designed to measure, based on people’s experience and available knowledge |
| Inter-rater reliability: Degree to which measurements are the same when obtained by different persons | Concurrent validity: Degree to which a measurement instrument produces the same results as another accepted or proven instrument that measures the same variable |
| Test-retest reliability: Degree to which the same test produces the same results when repeated under the same conditions | Predictive validity: Degree to which a measure accurately predicts expected outcomes |
| Equivalence reliability: Degree to which alternate forms of the same measurement instrument produce the same results | Construct validity: Degree to which a test measures the theoretical construct it intends to measure |

“end” may refer to changes in desired learner outcomes. Table 4 demonstrates 1 way to group studies using the proposed framework. Using this format, one can compare (for a given group of learners) the proportion of studies with positive outcomes for interventions with different intensities. Of note, tallying or “vote counting” results by studies with positive, negative, and inconclusive results is simplistic and may precipitate erroneous conclusions. Vote counting does not take into account sample sizes of individual studies and may miss positive outcomes if many studies are small or if the effect of the intervention is small.

Theoretically, content analysis could be used to help classify the components of educational interventions. However, content analysis is a method designed primarily for analyzing transcripts or other text. Relatively few educational studies provide adequately detailed descriptions of interventions to be used for content analysis.

In a review of the effectiveness of behavioral interventions to modify physical activity behaviors, Holtzman and colleagues had difficulty pooling studies because of diversity of outcomes, absence of critical information, and diversity of study populations and designs (48). However, the authors were able to calculate effect size (the difference between groups in a given measure divided by the standard deviation associated with that measure) (48). The Figure shows the effect size at last follow-up for each behavioral intervention at different levels of evaluation. Calculation of effect size enables the reviewer to compare effectiveness among interventions. The calculation and pooling of effect sizes can be useful when interventions use different outcome measures if the studies are significantly homogeneous otherwise (48). Calculation of effect size is an approach to synthesis that can be adapted to educational interventions that use different outcome measures.

Table 4. Grouping Studies To Demonstrate Positive Outcomes*

<table>
<thead>
<tr>
<th>Learner Type</th>
<th>Intervention</th>
<th>Intensity</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intervention</td>
<td>Duration, frequency</td>
<td>Knowledge</td>
</tr>
<tr>
<td>Learner type 1 (n = ...)</td>
<td>Intervention 1</td>
<td>Duration, frequency 1</td>
<td>P, N, I</td>
</tr>
<tr>
<td></td>
<td>Intervention 1</td>
<td>Duration, frequency 2</td>
<td>P, N, I</td>
</tr>
<tr>
<td></td>
<td>Intervention 2</td>
<td>Duration, frequency 1</td>
<td>P, N, I</td>
</tr>
<tr>
<td></td>
<td>Intervention 2</td>
<td>Duration, frequency 2</td>
<td>P, N, I</td>
</tr>
</tbody>
</table>

* I = inconclusive; N = negative; P = positive.

Figure. Example of use of effect size drawn from review of studies evaluating behavioral interventions to modify physical activity behaviors.

In contrast to qualitative reviews, quantitative reviews or meta-analyses rely on statistical methods to synthesize the evidence. Meta-analyses are more common for reviews of therapeutic clinical trials, diagnostic test evaluations, and epidemiologic studies than for reviews of educational interventions. Their statistical methods are used to determine the following: whether and the extent to which the results of studies are similar, the overall estimate of effectiveness and the precision of the estimate, and whether dissimilarities can be explained (49). However, the paucity of strong study designs and heterogeneity among studies generally limit reviewers’ ability to use quantitative approaches when synthesizing the evidence on educational interventions.

When study effect estimates are heterogeneous, meta-regression may be a useful analytic tool. Meta-regression is a statistical method for exploring possible sources of heterogeneity by evaluating the relationship between study-level covariates (for example, average age) and the effect estimate provided by each study. This approach is used when a statistical interaction exists between a study-level characteristic and the intervention of interest. Meta-regression can help to synthesize results of interventions with multiple components by evaluating the contribution of each component to the overall effect.

Meta-regression approaches include fixed effects, random effects, and control rate meta-regression, with the form of the model (for example, logistic or linear) dependent on the outcome measure being studied (50). The fixed-effects approach is appropriate if there is no source of variation in the estimated effect beyond differences in observed covariates (50). In random-effects meta-regression, a random study effect is modeled to account for between-study variation beyond that due to stochastic variation or observed covariate differences between studies. In control-rate meta-regression, a covariate is the outcome rate in the control group. This method is appropriate when the intervention effect depends on prevalence of the outcome in the control group; this prevalence is often a surrogate for other aspects of the underlying population (50). If this approach is not implemented properly, regression to the mean guarantees a strong relationship of the control rate with any effect measure that uses the control rate (for example, an odds ratio). These meta-regression methods may be particularly relevant when interventions and learner groups are heterogeneous.

Meta-regression has been used to assess the effect of individual components of interventions in health care (51, 52) and in the synthesis of educational interventions. For example, in a systematic review of the effect of diabetes patient education on glycemic control, Ellis and colleagues used meta-regression to determine which components of an intervention accounted for variance in glycemic control (53). They determined that face-to-face education, cognitive reframing, and exercise content accounted for 44% of the improvement in glycemic control (53).

While meta-regression is a valuable tool, it is susceptible to ecologic fallacy because the difference among studies based on a study-level factor may not mirror the relationship found within a study based on that factor measured at the individual level. Therefore, the resulting interpretation is more tenuous than if it were based on a within-study analysis. Also, meta-regression has limited power to explore the effects of multiple factors if the number of studies is small.

### Table 5. Recommendations for Improving the Reporting and Synthesizing of Educational Interventions

<table>
<thead>
<tr>
<th>For reviewers</th>
<th>For educational researchers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search databases likely to include educational interventions.</td>
<td>Submit key words that help educators find the report in computerized databases.</td>
</tr>
<tr>
<td>Be familiar with keywords and subject headings used in databases to improve yield.</td>
<td>Explain why chosen study design was the strongest possible with available resources.</td>
</tr>
<tr>
<td>Search targeted educational Web sites.</td>
<td>Describe baseline characteristics of targeted learners and comparison group(s) in detail and demonstrate that groups are comparable.</td>
</tr>
<tr>
<td>Describe search strategies, including terms, limitations, and exclusions.</td>
<td>Provide enough detail on educational content and teaching strategies to replicate.</td>
</tr>
<tr>
<td>Identify sources, including electronic databases, Web sites, hand searches, and expert consultation.</td>
<td>Describe context of intervention (setting, timing, targeted learners, and resources).</td>
</tr>
<tr>
<td>Report number of articles obtained and justify exclusions.</td>
<td>Use objective evaluation methods when possible.</td>
</tr>
<tr>
<td>Use established study design hierarchy to classify studies by design.</td>
<td>Report reliability and validity of instrument(s).</td>
</tr>
<tr>
<td>Identify confounding factors and differences between groups that bias results.</td>
<td>Acknowledge limitations of subjective evaluation methods used.</td>
</tr>
<tr>
<td>Acknowledge limitations of subjective evaluation methods used.</td>
<td>Assess reliability and validity of evaluation instruments and emphasize studies that used valid and reliable instruments.</td>
</tr>
<tr>
<td>Focus on objective evaluation methods.</td>
<td>Consider using evidence tables, summary tables, and figures to display results.</td>
</tr>
<tr>
<td>Consider using numeric scales and evidence grading to assess study quality.</td>
<td>Consider using quantitative techniques, including effect size and meta-regression if appropriate.</td>
</tr>
</tbody>
</table>

### Conclusion and Recommendations

We have described methodologic challenges to reporting and reviewing interventions in medical education. To address these challenges, we’ve provided recommendations for educational researchers and reviewers in Table 5. Specifically, we recommend that educational researchers increase the rigor with which they design, analyze, and report educational interventions. High-quality studies and unam-
Challenges in Systematic Reviews of Educational Intervention Studies

26. Thomas PA, Kern DE. Internet resources for curriculum development in...
34. Prideaux D. Researching the outcomes of educational interventions: a matter of design. RTCs have important limitations in evaluating educational interventions [Editorial]. BMJ. 2002;324:126-7. [PMID: 11799017]
Systematic reviews are a type of literature review that uses systematic methods to collect secondary data, critically appraise research studies, and synthesize findings qualitatively or quantitatively. Systematic reviews formulate research questions that are broad or narrow in scope, and identify and synthesize studies that directly relate to the systematic review question. They are designed to provide a complete, exhaustive summary of current evidence relevant to a research question. For example